Learned ontology guided opinions analysis of extracted aspects from online product reviews

Ravi kumar K.C.[1,*], Teja Santosh Dandibhotla[2], Vishnu Vardhan Bulusu[3]

1. JNTUH, Telangana State, India

2. Sreenidhi Institute of Science and Technology
Yamnampet, Ghatkesar,Telangana State, India

3. JNTUHCEM, Manthani, Telangana State, India

kcravikumar1971@gmail.com

ABSTRACT. *The opinions expressed by the consumers on online product reviews in e-commerce websites play major role in judging the evaluative character of the product aspect. These expressed opinions lack conceptual preciseness allowing consumers to use them in both syntactically and semantically different ways (lexical variations) on various aspects in the reviews. Also some section of consumers present their opinions in the implicit manner. The evaluation of these types of opinions for opinion orientations raises the semantic gap between the human language and the actual opinionated knowledge. Thus, extracting all these types of opinions on the product aspects may bridge the semantic gap and thereby improving the accuracy of the opinion orientation. In this paper, iterative ontology learning approach is carried out in order to solve the aforementioned problems. In the proposed method, first the pre-processed product reviews are analyzed for extracting opinionated lexical variations. Then, the reviews are further analyzed to extract the implicit opinions. Further, these opinionated lexical variations and implicit opinions with the reviews are formalized for ontology learning. The aspect, opinion pair is formed by reasoning the learned ontology. Finally, the aspect's opinion orientation is ascertained by using the sentiwordnet scores in the improved geodesic distance metric. The evaluation of semantic orientation of opinions using the learned ontology guidance against the state-of-the-art approaches shows the effectiveness of the proposed method.*

RÉSUMÉ. *Les opinions exprimées par les consommateurs sur les évaluations de produits en ligne dans les sites Web de commerce électronique jouent un rôle majeur dans l'appréciation du caractère évaluatif de l'aspect produit. Ces opinions exprimées manquent de précision conceptuelle, ce qui permet aux consommateurs de les utiliser de manières différentes du point de vue syntaxique et sémantique (variations lexicales) sur divers aspects des appréciations. De plus, certains consommateurs présentent leurs opinions de manière implicite. L'évaluation de ces types d'opinions pour les orientations d'opinion soulève le fossé* *sémantique existant entre le langage humain et le savoir réellement exprimé. Ainsi, extraire tous ces types d'opinions sur les aspects du produit peut combler le vide sémantique et ainsi améliorer la précision de l'orientation des opinions. Ainsi, extraire tous ces types d'opinions sur les aspects du produit peut combler le fossé sémantique et ainsi améliorer la précision de l'orientation des opinions. Dans cet article, une approche d'apprentissage ontologiques itératives est réalisée afin de résoudre les problèmes susmentionnés. Dans la méthode proposée, l'appréciation de produits pré-traités sont d'abord analysées pour extraire les variations lexicales jugées. Les appréciations sont ensuite analysées pour extraire les opinions implicites. De plus, ces variations lexicales et les opinions implicites avec les revues sont formalisées pour l'apprentissage ontologique. L'aspect de la paire d'opinions est formé en raisonnant l'ontologie apprise. Enfin, l'orientation des opinions de l'aspect est déterminée en utilisant les scores de sentiwordnet dans la métrique de distance géodésique améliorée. L'évaluation de l'orientation sémantique des opinions à l'aide du guide d'ontologie appris par rapport aux approches d'état de l'art montre l'efficacité de la méthode proposée.*

KEYWORDS: *online reviews, product aspects, opinions, adjective, lexical variations, implicit opinions, ontology learning, semantic orientation.*

MOTS-CLÉS: *appréciations en ligne, aspects produits, opinions, adjectif, variations lexicales, opinions implicites, apprentissage ontologique, orientation sémantique.*

DOI:10.3166/JESA.51.25-49 © 2018 Lavoisier

1. Introduction

Over the last two decades, the amount of web information has exploded in a rapid way in all formats of data. The online flow of information is on a constant rise. The web content is growing at a lightning fast speed. The need for sharing information among the web users is also increased. The changes in the web usage patterns have led to an immense communication and social interaction among the internet community. The term web 2.0 is coined by Tim O' Reilly has this social web as a part of it. Popular web 2.0 online shopping websites like Amazon, Flipkart and etc., which utilize E-commerce Business-to-Consumer (B2C) business model for conducting online transactions are contributing to the content development over the web. This is by providing information on large amount of goods and services. They are providing the customers to write their reviews on the already purchased products that are offered from their website.

The customers who are in need of purchasing a product visit the online shopping website, search for the relevant information and evaluate the product using the online reviews. The product evaluation is performed by learning the semantic orientation of the opinion word. The semantic orientation of the word indicates the tendency of the word deviating from its actual norm for its semantic group (Hatzivassiloglou and McKeown, 1997). This tendency of the opinion word is analyzed by calculating the orientation of the adjectives by collecting the polarized information from the reviews.

An automated system for measuring semantic orientation has successful applications in text classification, tracking opinions in online discussions and analysis of survey responses. In the context of online product reviews customers write the adjectives (Pang *et al*., 2002) to convey the opinionated information. These adjectives are analyzed for the semantic orientation for the corresponding aspects. The adjective-adverb combination (Benamara *et al*., 2007) is also used to pen the opinionated reviews. The opinion word retrieval accuracy is good as it is evaluated with adjectives.

The knowledge expressed in the customer reviews on the product for opinions by using natural language is often vague and underspecified (Buitelaar and Cimiano, 2008). The implicit opinionated sentences (Huang *et al*., 2017) and the lexical variations (Grondelaers *et al*., 2012) used in writing the reviews increase this semantic gap in terms of expressing opinionated information on the products and their aspects. The research goal of this work is to bridge the aforesaid semantic gap and improving the accuracy of the opinion orientation of the aspect in the e-commerce product reviews analysis.

In order to carry out this research goal, the lexical variations and the implicit opinions are extracted on the aspects from the product reviews and are formalized in the form of ontology. This paradigm of ontology engineering is called as ontology learning from text (Buitelaar *et al*., 2005). This learned ontology is iterative in manner. Then the aspect, opinion word pair is formed by reasoning this learned ontology. Finally, the aspect's opinion orientation is carried out by using the sentiwordnet scores in the modified geodesic distance metric (Ravi Kumar *et al*., 2017). Finally, the accuracy of the improved opinion word is estimated and the semantic orientation is presented.

This paper is organised as follows: related work is described in section 2, the proposed approach is explained in Section 3, the experimental procedure of the proposed approach and the results of opinion orientation of the proposed method compared with the existing approaches is discussed in Section 4 and finally the conclusions and the scope for future work are presented in Section 5.

2. Background and related works

The E-Commerce platforms allow the consumers of the product to pen their feelings in the form of online reviews using natural language. More often the reviews are written in free flow, unstructured format allowing reviewers to write lengthy reviews. These expressed writings inVolve in them the knowledge levels of the language of the reviewer in the form of sentences. The automated understanding of human intentions from these review sentences for a fellow human is easy. The same task is very challenging to carry out by the machine. In order to mitigate this problem, a popular tool namely Natural Language Processing (NLP) is used. NLP provides the ability to the machine to analyze the human language (either speech or text) and get the understanding of the language with the maximum accuracy.

In order to carry out this task by the machine, the processing environment depends on various dictionaries and lexicons. A lexicon is a collection of lexemes (basic unit of language with one or several words intended to convey the meaning as a whole) in the alphabetical order. WordNet is a lexical knowledge base for English language. It groups the English words into sets of synonyms called synsets. The major purpose of WordNet is to support automatic text analysis in many of the artificial intelligence applications. Most of the synsets are connected to other synsets in the WordNet through semantic relations. These relations differ based on the type of word. The various semantic relations based on the type of the word are: I) hypernyms, hyponyms, holonyms and meronyms fall under Noun category, II) hypernym, troponym, entailment, coordinate terms fall under Verb category, III) related nouns, similar to, participle of verb fall under Adjective category and iv) root adjectives fall under adverbs category. These semantic relations are used for determining similarity between the concepts. SentiWordNet is an enhanced lexical resource which contains sentiment scores for the WordNet word types. The main purpose of SentiWordNet is to support the task of opinion mining.

The lexical variation is a NLP component which has got major research attention over more than two decades. Much natural language processing research implicitly assumes that word meanings are fixed in a language community, but in fact there is good evidence (Johnson, 2002) that different people probably associate slightly different meanings with words. This is understood by decomposing the lexeme details into three parts namely;

(I) Morphological information

(II) Syntactic information

(III) Semantic information

The morphological information provides the information on lexeme composition. The syntactic information of a lexeme provides the word Part of Speech (POS) and the word variation in terms of suffix or prefix information. The semantic information disambiguates the context of the word by providing the concept of the term with its gloss. The semantic information also provides the synonyms and antonyms associated with the lexeme.

Dirk Geeraerts *et al*. (2012) explained varieties of lexical variations and their influence on contextual variation. The authors restricted their explanation to the fashion domain and the terms usage in British English and American English languages. Simonetta Montemagni and Martijn Wieling (2016) tracked linguistic features underlying lexical variation patterns from Tuscany region (geographical location clusters) dialects using Concept-Lexicalization pairs. These works emphasize the important information of speaker words association with the country or region specific intelligibility.

Motivated with the above research Svitlana Volkova *et al*. (2013) explored the lexical variations based on gender and age demographics to improve multilingual sentiment analysis in social media. The researchers have analyzed the subjective sentences in English, Spanish and Russian Twitter data. Jantima Polpinij and Aditya K. Ghose (2008) classified sentiments of the online product reviews using the engineered lexical variation ontology. The concepts in this ontology are annotated with the extracted words after performing morphological and statistical analysis on the extracted tokens from the unstructured online reviews.

The topic of implicit opinions is pointed out in the year 2012. Least percentage of online reviews was expressed in terms of the opinion on the product or aspect in the indirect manner. These opinions are called as implicit opinions. Bing Liu (2012) critically analyzed about the issue of implicit opinions towards valuable opinion mining. Kumar Ravi and Vadlamani Ravi specified (2015) the importance of implicit opinions in improving the precision of the extracted product aspects. Recently Huang *et al*. (2017) extracted the implicit opinion words from the implicit opinion snippets and clauses and assigned the polarity class label. The classifiers used to carry out this task are Support Vector Machines (SVM) and ConVolutional Neural Networks (CNN). It has been observed that CNN significantly outperformed SVM on implicit opinion snippets and clauses.

The research on ontology learning applied to social media content is less studied. Lau *et al*. (2009) developed light weight fuzzy domain ontology learning method to automatically generate concept hierarchies on the extracted text from online forums. Tho *et al*. (2006) applied fuzzy formal concept analysis to build domain ontology automatically. Recently Ahmad *et al*. (2016) proposed an integrated text mining system for electronic products ontology learning and sentiment analysis. The average precision across four machine learning classifiers for ontology learned sentiment classification is 92.7% on training dataset and the average precision of 85.5% on test dataset.

The semantic orientation of texts is an age old classical work for more than five decades. Osgood *et al*. (1958) identified several pairs of bipolar adjectives that greatly influence the shift in the orientation of the opinion words. Hatzivassiloglou and McKeown attempted (1997) to predict the orientation of the opinion words by analyzing the pairs of adjectives bounded by conjunctions. Turney and Littmann (2003) approached the problem by using a seed set to bootstrap the process of opinion word identification for the first time. Once the opinion words are identified, Pointwise Mutual Information (PMI) was calculated on the identified opinion word and the term in the seed set. The work on determining the orientation of the terms is concentrated on pairs of adjectives bounded by conjunctions. The researchers considered only 657/679 documents (labeled Positive/Negative) in which the adjectives bound by conjunctions are available from the Wall Street Journal (WSJ) corpus (Hatzivassiloglou and McKeown, 1997).

Kamps *et al*. (2004) focused on the relations between the words defined in the WordNet (Fellbaum, 1998). They calculated the relative distance from the two seed terms to the identified opinion word to determine the orientation of the opinion word. The work of solving the ambiguity of the terms that appear in both the Positive and Negative categories was never concentrated. They removed those terms from the sets and experimented on the reduced sets. The number of considered terms after removing the ambiguous entries is 1614/1982. They restricted the adjectives in the analysis to 663 from the total 3596 terms of Turney and Littmann as used in Turney and Littman (2003). This is because the synonymy relation graph of WordNet evaluates only those adjectives that are in the path of the graph bounded with the seed terms at the ends of the graph. Recently, Emiel van Miltenburg (2016) calculated the distance between two adjectives by obtaining derivationally related forms of the adjectives. These derivationally related forms are associated with the adjective lemmas.

In this body of literature, the following shortcomings are identified: First, the works on lexical variations never concentrated on irregular verbs, irregular adverbs and irregular adjectives in the opinions analysis. Further, the works on implicit opinions never concentrated on assigning explicit opinion words to the corresponding aspects extracted from implicit opinion clauses by carrying out the analysis on the extracted explicit opinionated aspects. Furthermore, all the works except Ahmad *et al*. work learnt a fuzzy ontology from the text which is helpful for representing and reasoning the uncertainty in the domain knowledge. The non-fuzzy ontology learning is to be carried out as the domain under analysis is clear and certain. Finally, the geodesic distance measure is to be modified in such a way that the relative distance which is calculated to determine the semantic orientation uses the Sentiwordnet 3.0 scores.

3. Methodology

The determination of semantic orientation of opinion words using sentiwordnet scores is presented in Figure 1. Input to the model is the online reviews**.** Initially, the incoming product reviews are pre-processed. The steps in pre-processing are namely review tokenization, stop words removal and POS tagging. The process of review tokenization divides the sentence into individual tokens. Then, the stop words list is applied on the tokens to remove those words which carry no meaning in the analysis. The stop words are compiled from the reviews itself. This compilation is carried out by sorting the terms in the decreasing order of collection frequency and thereby hand-filtering those terms for their semantic content relative to the domain of the product reviews. However, the negations, conjunctions and interjections are not removed. Finally, POS tagging is carried out on the list of filtered tokens to associate the unambiguous word categories with each of the token. The Stanford log-linear Part of Speech tagger is used for tagging the tokens (Toutanova and Manning, 2000).

In identifying and extracting opinionated lexical variations from the online reviews three PoS tagged words with their tense and person position variants are considered. These are Verbs (VB) and their variants, Adverbs (RB) and their variants, and adjectives (JJ) and their variants. These PoS tags and their variants are most likely to reveal customers emotions and attitudes on the product and their aspects. The irregular verbs list, irregular adverbs list and irregular adjectives list are collected from the various websites. These are listed after the 6th section below.

The POS tagged words and their variants from the online reviews are checked against these lists with the above mentioned WordNet options. When a match is found it is stored in the list of opinionated lexical variations. In order to extract the implicit opinion words following steps are carried out. First, the implicit opinion sentences are identified. Then the aspects present in these sentences are extracted using nouns. Further, the explicit opinion words (adjectives) written for the extracted aspects in the explicit opinionated reviews are analyzed for synonym grouping. The present research work considers explicit aspects for the opinions analysis.

The assignment of the opinion word is performed by carefully mapping the polarity of the implicit opinion clause with the polarity of the explicit opinionated reviews on the basis of aspects. Finally, the aspect-opinion ontology learning is carried out to extract maximum aspect-opinion word pairs from the reviews.

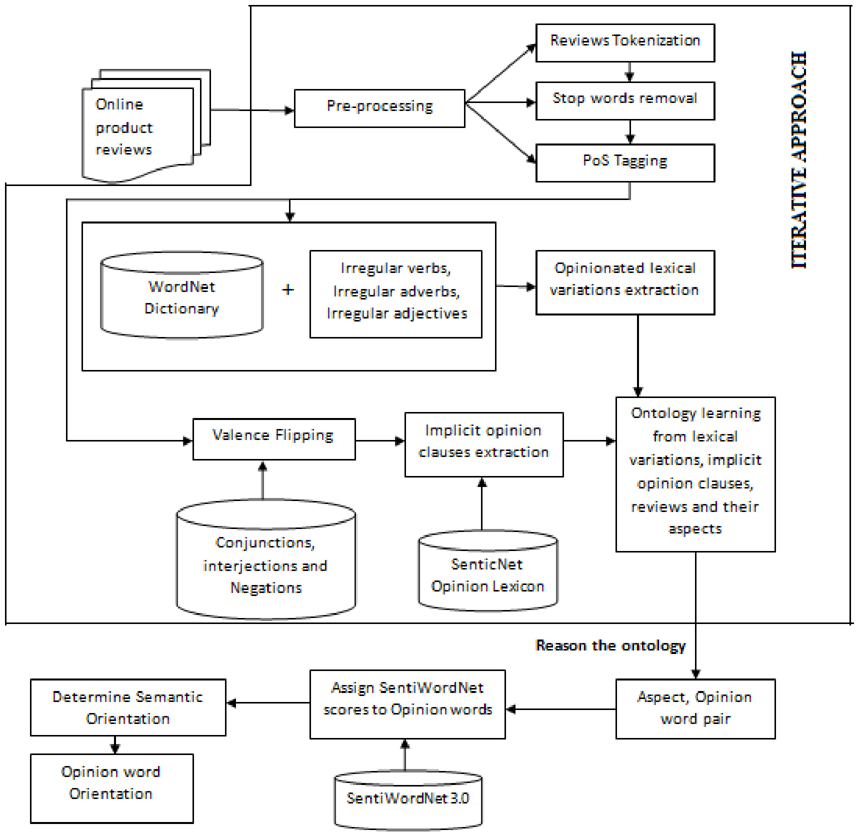


Figure 1. Proposed model

3.1. Ontology learning from online product reviews by including opinionated lexical variations and implicit opinion words in addition to adjectives

In order to improve the opinion word retrieval accuracy on the product aspects the extracted lexical variations and the implicit opinions in addition to adjectives are used as input terms to start the ontology learning process. The learned ontology is iterative in manner. In this way the processing of a new batch of product reviews (e.g., from current day) inVolves reusing the earlier learned ontology (e.g., a day before) which is updated further. The aspect, opinion word pair is formed finally by reasoning the ontology. The pipeline diagram of the ontology learning from product reviews is presented in Figure 2 below.

1. Product domain specific termsand synonyms extraction	2. Concept learning for the extracted terms and synonyms	3. Taxonomy learning among the learned concepts	4. Non-taxonomic Relations Learning among the concepts	5. Learning of axioms and relation characteristics	6. Semantic Annotation of terms with relation based concepts

Figure 2. Pipeline diagram of the ontology learning from product reviews

The aspect-opinion ontology learning is carried out by the following steps.

(1) Extracting domain specific terms and their synonyms from product reviews

The domain for which the terms and synonyms are extracted is online electronic product reviews. The product name and the aspects are extracted from the reviews. The nouns are identified as domain specific terms. Various linguistic patterns from review sentences are identified in order to extract the terms. These patterns are as follows

Linguistic Pattern1: term_NN term_JJ

The noun term is extracted as an aspect when it is found in combination with the adjective as suffix term in the review sentence. This adjective was identified as opinion word earlier.

Linguistic Pattern2: term_NN term_JJ **and** term_JJ *nsubj* term_NN

The noun term is extracted as an aspect when it is found in combination with the adjective term in the review sentence. Also the noun term becomes the syntactic subject for the governed adjective.

Linguistic Pattern3: term_JJ term_NN

The noun term is extracted as an aspect when it is found in combination with the adjective as prefix term in the review sentence. This adjective was identified as opinion word earlier.

Linguistic Pattern4: term_JJ term_NN **and** term_NN *dobj* term_JJ

The noun term is extracted as an aspect when it is found in combination with the adjective term in the review sentence and the noun term becomes the object of the adjective.

Linguistic Pattern5: term_JJ term_NN **and** term_NN *amod* term_JJ

The noun term is extracted as an aspect when it is found in combination with the adjective term in the review sentence and the adjective term modifies the meaning of the noun term.

Linguistic Pattern6: term1_NN term2_NN term3_JJ **and** term1_NN *compound* term2_NN

The second noun term is extracted as an aspect from the review sentence when it is found in combination with the first noun term as prefix and adjective as suffix. The first noun term serves to modify the second noun term.

Linguistic Pattern7: term1_NN term2_NN term3_JJ **and** term1_NN *conj* term2_NN

The two noun terms are extracted as aspects when they are found in combination with the adjective as suffix. Also in this review sentence, the first noun term relates to the second noun term using a conjunction (and, or, but) relation.

Linguistic Pattern8: term1_NN term2_RB term3_JJ **and** term3_JJ *nsubj* term1_NN **and** term3_JJ *advmod* term2_RB

The noun term is extracted as an aspect when it is found in combination with the consecutive adverb and adjective terms. Also in this review sentence, the noun term becomes the syntactic subject for the governed adjective and the adverb term serves to modify the meaning of the adjective term.

Linguistic Pattern9: term_VB term_NN **and** term_VB *dobj* term_NN

The noun term is extracted as an aspect when it is found in combination with the verb term in the review sentence and the verb term becomes the object of the noun term.

Linguistic Pattern10: term_NN/term_NNS

The noun term or the plural noun term is extracted as an aspect when it is the only term in the review sentence.

The task of identifying and grouping synonyms for the extracted terms is viewed as a word sense disambiguation (Wang and Hirst, 2012). The domain dependent synsets are to be formed. In order to group synonyms from the reviews, three principles are applied. These are minimality, coverage and replaceability.

(I) Minimality: The minimum set of associated words that distinctively identifies the meaning is first used to create the synset. The associated words and the gloss for these words are obtained from the Google dictionary to disambiguate the intuition of sense. For example, to create the synset for the review words image and picture to the extracted term photo, first the gloss is obtained. Then the words are checked for associativity. Now these two words are created as synset. The synset is {image, picture}.

(II) Coverage: Next, the synset should contain all the words indicating a particular meaning. The words are written down in the decreasing frequency of their occurrence in the reviews. For example for the extracted term photo, the review word snapshot is disambiguated based on its gloss. Also this word is less frequent from the reviews. The word is then added to the synset. Now the synset is {image, picture, snapshot}.

(III) Replaceability: The words forming the synset should be mutually replaceable in a specific context. For example in the following phrases, great image and good snap, image and snap can replace each other.

(2) Concept Learning for the extracted terms and synonyms

The higher level concepts for the extracted terms are learned using the ConceptNet multilingual knowledge graph. For example, the term photo has the higher level concept graphic (Speer and Havasi, 2012). The concept is finalized based on the synonyms grouped as synset in the previous step and related terms from the ConceptNet.

(3) Taxonomy Learning among the learned concepts

The super-concept and sub-concept relational hierarchy is learned towards formal representation of ontology. The Word Class Lattice (Navigli and Velardi, 2010) (WCL) dataset is used to arrange the super-concept and sub-concept in the form of taxonomy.

(4) Semantic (non-taxonomic) Relations Learning among the concepts

The non-taxonomic relations among the concepts are learnt in order to have clear knowledge model. In order to learn non-taxonomic relations among the concepts the frequent associations between the concepts is obtained from the reviews. The label is then provided to these concept associations.

(5) Learning of axioms and relation characteristics

The axiom disjointness between the concepts is learned from the conjunctions among the corresponding terms from the reviews. The symmetric relation between the concepts is learned from the reviews based on the co-occurrence of the POS patterns in both directions. For example, the learnt relation isExtractedBasedOn is symmetric as the PoS patterns Term_NN Term_JJ and Term_JJ Term_NN are analyzed for double propagation towards aspects and opinion words identification (Qiu *et al*., 2011).

(6) Semantic Annotation of terms with relation based concepts – Ontology Population

The extracted terms are finally annotated under the learned concepts and relationships. Also the synonyms are also placed under the concepts. After carrying out all the steps in the ontology learning from reviews, the final ontology learned is illustrated in Figure 3. below.

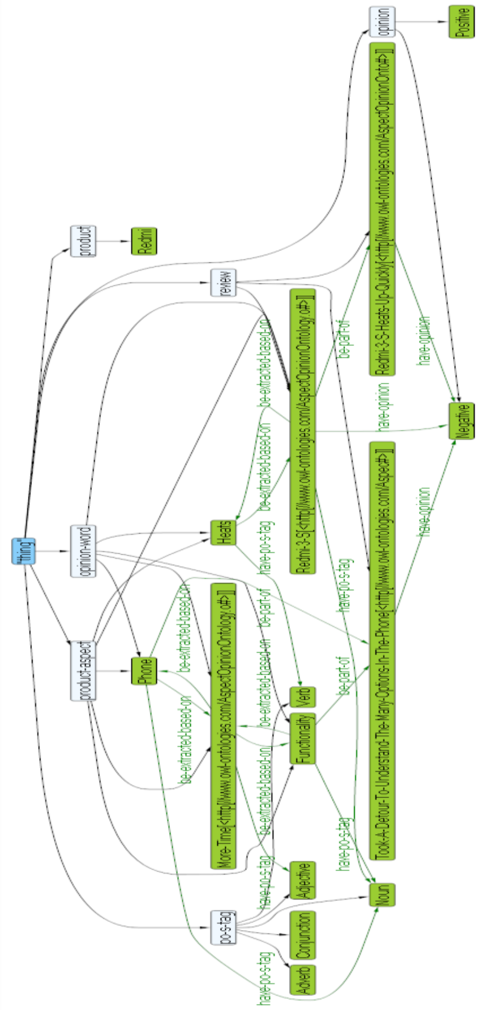
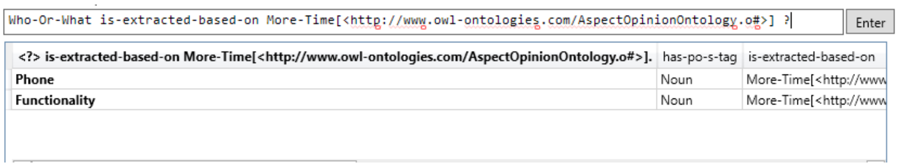


Figure 3. Learned aspect-opinion ontology

Reasoning the AOO for aspect, opinion word pair is presented in Figure 4. below.



Who-Or-What is-extracted-based-on More-Time[<http://www.owl-ontologies.com/AspectOpinionOntology.o#>] ?

Enter

<?> is-extracted-based-on More-Time[<http://www.owl-ontologies.com/AspectOpinionOntology.o#>].	has-po-s-tag	is-extracted-based-on
Phone	Noun	More-Time[<http://www
Functionality	Noun	More-Time[<http://www

Figure 4. Reasoning the AOO for aspect, opinion pair

3.2. Semantic orientation of aspect specific opinion words using sentiwordnet scores in the modified geodesic distance metric

The output from the learned ontology is the aspect, opinion word pair. Now the orientation of the opinion word is to be calculated. The traditional geodesic distance metric is modified in order to determine the opinion orientation of every opinion word from the corpus. The variations in the method for determining the orientation of an opinionated term using the opinion word senses and the Sentiwordnet scores were presented. The process is composed of following steps:

1. A standard opinion lexicon in which two sets of adjectives are present is considered as input for bootstrapping. These sets represent two categories namely Positive and Negative. Two seed terms ‘good’ and ‘bad’ represent the two categories are taken into consideration.

2. The sizes of the Positive and Negative adjective sets are increased by adding the synonyms and antonyms of the adjectives using WordNet.

3. The increased sizes of Positive and Negative adjective sets are used to compare with the obtained adjectives from the dataset. Once the dataset adjectives are matched with the opinion lexicon adjectives, then the dataset adjectives are considered as opinion words. This completes the identification of opinion words from the dataset.

4. The opinion word and the seed terms are assigned with the numerical scores available under adjective category from Sentiwordnet. This is carried out by finding the contextual clues surrounding the opinion word. These contextual clues will help to disambiguate the sense of the opinion word. The contextual clues are finalized based on the typed dependency grammatical relations.

5. The distance between the opinion word and the seed term and the distance between the seed terms is calculated as given below.

$$\text{distance}(w_i,w_j)=\text{sentiwordnetscore}(w_i)\text{-sentiwordnetscore}(w_j) \qquad (1)$$

where w_i is either the opinion word or the seed term and w_j is the seed term. The distance measure is modified as the application of distance is carried on nonhierarchical semantic network i.e., on adjectives (Sun *et al.*, 2017).

6. The semantic orientation (SO) of the opinion word is determined as given below.

$$SO(\textit{opinion word})=\frac{\textit{distance}(\textit{opinion word, bad})-\textit{distance}(\textit{opinion word, good})}{\textit{distance}(\textit{good, bad})} \qquad (2)$$

7. The opinion word is deemed to be positive when the orientation measurement is greater than zero, and negative otherwise.

Step 2 is based on the premise that the lexical relations used in this expansion task which define a relation of orientation. It is possible that two synonyms may have same orientation and two antonyms have opposite orientation. In step 4, the basic assumption is that the terms with a similar orientation tend to have similar glossaries. The similarity or difference between the opinion word and the seed term is based on identifying the appropriate senses in the context in which the opinion word is written in the document. The senses of the seed term is going to change based on the context of the opinion word under analysis. The replacement with the sentiwordnet scores in step 5 enables to determine the orientation of any opinion word with the help of SO measure specified in step 6.

4. Experimental setup and data analysis

The laboratory environment that was setup for carrying out learned ontology guided opinions analysis experiment typically consists of a dedicated computer machine running with Windows 8 operating system with 1 TB hard disk as secondary memory. The capacity of the primary memory is 4GB. The technologies used in this experiment to implement the model are namely WordNet dictionary, SenticNet and SentiWordNet opinion lexicons, ConceptNet knowledge graph and Artificial Intelligence based Ontology which is learned.

The datasets used for this task are the collection of five categories of product reviews from Amazon. GPS devices, Tablets, Laptops, Smart phones and cameras are the product categories for which the reviews are considered for analysis. In each product category, 100 products are considered from the E-commerce application. Table 1. presents the details of the datasets used for this experiment.

Table 1. Dataset details

Document attributes	Values
Number of review documents	167458
Minimum sentences per review	1
Maximum sentences per review	43
Average number of reviews written by customers	5.78
Average number of reviews written on the product	48.47

The pre-processing of data is carried out by removing stop words and non English words. PoS tagging is performed on the obtained set of words. The adjectives, verbs and adverbs present in the review sentences are analyzed for opinion words. The WordNet dictionary provides very important pieces of information namely synonyms, antonyms, hypernyms, troponyms, derivationally related forms and coordinate terms for a verb. For an adverb WordNet provides synonyms and stem adjectives. For an adjective WordNet provides synonyms, antonyms, sense information and derivationally related forms. These details provide the morphological, syntactic and semantic information on the lexeme. The lexeme details for the verb “block” is tabulated in Table 2 below.

Table 2. Lexeme details for the verb “block”

Lexeme	Block	#Word
Morphological information	Single	#Word formation – Single word
Syntactic information	PoS Tag {VB} Suffix {s, ed}	#Word class #Word variations
Semantic information	Hypernym Concept {prevent} Synonyms {obstruct, stop, hinder} Antonyms { unblock, unfreeze, free, release}	#Concept of the given word #Words that are similar in meaning of the given word #Words that have opposite meaning of the given word

The corpus statistics of irregular verbs, irregular adverbs and irregular adjectives are presented in Table 3 below.

Table 3. Corpus statistics of irregular verbs, irregular adverbs and irregular adjectives

Irregular verbs	Irregular adverbs	Irregular adjectives
21765	16888	38596

The statistics on the number of irregular verbs, irregular adverbs and irregular adjectives that are considered for the comparison with the corpus statistics as mentioned earlier in Table 2 are provided in Table 4 below.

Table 4. Irregular verbs, Irregular adverbs and Irregular adjectives statistics for comparison collected from web

Irregular verbs	Irregular adverbs	Irregular adjectives
139	10	16

The impact of irregular verbs on opinion orientation of the product aspects is carried out on the datasets to find out whether irregular verbs have a significant influence on the opinion analysis. The number of aspect, opinion pairs identified using irregular verbs and the positive or negative implications of these pairs are tabulated in Table 5 below.

Table 5. Number of aspects, opinion pairs and their orientations from irregular verbs

Product category/No. of aspect, opinion pair identified using irregular verbs	Positive/Negative Impact	Total Number of irregular verbs	% of implied positive/negative opinions
GPS Devices/386	199+/187-	3588	5.5% / 5.2%
Tablets/678	358+/320-	4693	7.6% / 6.8%
Laptops/779	436+/343-	4036	10.8% / 8.4%
Smart phones/592	345+/247-	5121	6.7% / 4.8%
Cameras/756	447+/309-	4327	10.3% / 7.1%

The percentage of implied positive opinions and negative opinions identified by irregular verbs tabulated above provides the information that irregular verb has good influence on the opinion analysis.

The impact of irregular adverbs on opinion orientation of the product aspects is carried out on the datasets to find out whether irregular adverbs have a significant influence on the opinion analysis. The number of aspect, opinion pairs identified using adverbs and the positive or negative implications of these pairs are tabulated in Table 6 below.

Table 6. Number of aspects, opinion pairs and their semantic orientations from irregular adverbs

Product category/No. of aspect, opinion pair identified using irregular adverbs	Positive/Negative Impact	Total Number of irregular adverbs	% of implied positive/negative opinions
GPS Devices/412	396+/16-	2679	14.7% / 0.6%
Tablets/693	483+/210-	3954	12.2% / 5.3%
Laptops/806	669+/137-	2027	33.0% / 6.7%
Smart phones/624	452+/172-	4376	10.3% / 3.9%
Cameras/786	704+/82-	3852	18.2% / 2.1%

The percentage of implied positive opinions and negative opinions identified by irregular verbs tabulated above provides the information that irregular verb has good influence on the opinion analysis.

The negated words, conjunctions and interjections in the online reviews are analyzed carefully in order to change or flip the valence of the word. The negated words, conjunctions and interjections are compiled from the reviews itself. Finally, the opinion clauses that are manually annotated on the online reviews are analyzed. The statistics on the number of explicit opinionated clauses and implicit opinionated clauses from the total number of considered reviews are given below in Table 7.

Table 7. Statistics of clause extraction from the reviews corpus

Polarity	Positive	Negative	Total
Explicit opinionated clauses	67,321	21,078	88,399
Implicit opinionated clauses	40,372	15,412	55,784
Total clauses	1,07,693	36,490	1,44,183

Sample positive and negative implicit opinionated clauses manually annotated with polarity from online reviews are presented below.

- There are many hitches in this smartphone: Negative
- Took a detour to understand the many options in the phone: Negative
- The chrome app stops after entering a website: Negative
- Two minutes to edit the picture with refocus mode: Positive
- Most game apps do not curtail the mobile RAM performance: Positive
- This smartphone stores many pictures: Positive

The above presented sample manually annotated implicit opinion clauses are considered from high degree of annotation agreement result between the two human annotators. The two human annotators are experts who have familiar knowledge with E-commerce products and reviews domain. In order to evaluate the accuracy of the annotated dataset, Kappa Coefficient (k) was considered to compare the result of each annotator. The formula of Kappa Coefficient is as follows:

$$k = \frac{\Pr obability(observedagreement) - \Pr obability(chanceagreement)}{1 - \Pr obability(chanceagreement)} \tag{3}$$

where $Pr\,o\,bability\,(observedagreement)$ is the proportion that the two human annotators agree on a common annotation and $Pr\,o\,bability\,(chanceagreement)$ is the proportion that the two human annotators are expected to agree by chance. The obtained value of ‘k’ is 0.87. This result specifies that the two human annotators reached a high agreement in the implicit opinion clause annotation task.

After a careful analysis of the considered online product reviews it is observed that the opinion words are expressed explicitly when the consumers convey their views in the subjective manner. Whilst the implicit opinions are those which the consumers narrate their experiences in writing. Also, it is observed that the narrated writings are often vague and contain no opinion words on the target entity being described. This makes the task of opinion mining more challenging. The implicit opinion clauses of the review sentences provide the opinion on the aspect or entity in indirect manner at sentence level. The implicit opinion words are the opinions that are implied on the aspects which are written in the implicit opinion clause.

An analysis is carried out on these explicit opinion words in order to assign these opinion words to the corresponding aspects extracted from implicit opinion clauses. Various parameters for the number of explicit opinion words with synonym grouping and the number of identified aspects from implicit clauses is given below in Table 8.

Table 8. Explicit opinion words and Identified aspects from implicit clauses details

Parameter	Value/Percentage
Total number of explicit opinion words (without synonym grouping)	38596
Number of identified aspects from implicit clauses	29621
Total number of aspects from the reviews corpus	36724
Total number of explicit opinion words (with Synonym grouping)	37928
Percentage of identified aspects from implicit clauses	80.6%
Percentage of explicit opinion words which do not have any synonyms	1.7%

It is observed from the above table that majority of the implicit clauses available from the online reviews contain explicit mentions of the aspects. Also it is learned that very less percentage of explicit mentions of opinion words that are written in the online reviews have no synonyms.

The learned ontology is not fuzzy as the concepts and relationships learned are clear. The evaluation of the learned ontology is carried out by considering the Product Review Opinion Ontology (Santosh and Vardhan, 2016) (PROO) as the reference ontology. PROO ontology is a general ontology engineered for the smartphones domain. This ontology is general enough to assign instances from any of the electronic product reviews. The visualization of the learned ontology and the referred PROO ontology are illustrated below in Figure 5.

The PROO ontology is termed as O_R and the learned ontology is termed as O_L. C_L is the set of concepts in learned ontology and C_R is the set of concepts in the reference ontology. In order to find out whether the learned ontology is highly similar with the reference ontology various measures are computed. These are Lexical Precision (LP) and Lexical Recall (LR) and Taxonomic Precision (TP) and Recall (TR) based on common semantic cotopy (csc) for understanding about the common concepts in both concept hierarchies.

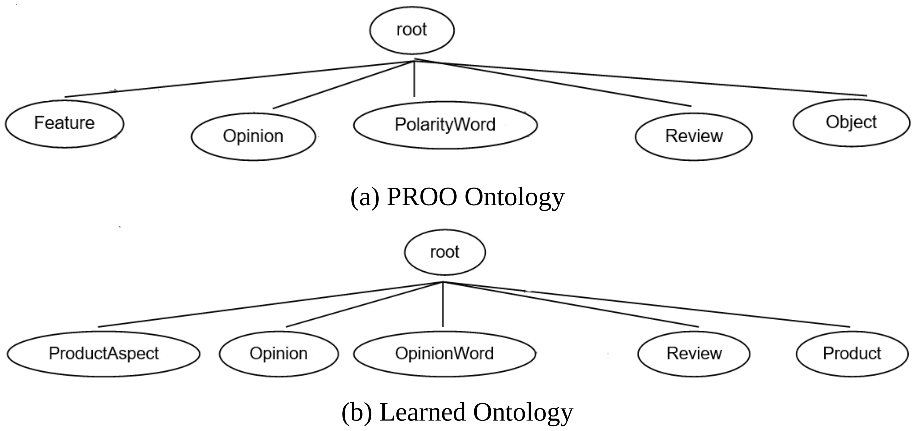


Figure 5. Visualization of reference ontology and the computed ontology

The formula for Lexical Precision (LP) and Lexical Recall (LR) is;

$$LP(O_L, O_R) = \frac{|C_L \cap C_R|}{|C_L|} \tag{4}$$

$$LR(O_L, O_R) = \frac{|C_L \cap C_R|}{|C_R|} \tag{5}$$

The formula for Taxonomic Precision (TP_{csc}) and Taxonomic Recall (TR_{csc}) is;

$$TP_{CSC}(O_L, O_R) = \frac{1}{|C_L \cap C_R|} \sum tp_{csc}(c_1, c_2, O_L, O_R) \tag{6}$$

Where

$$tp_{csc}(c_1, c_2, O_L, O_R) = \frac{|ce(c_1, O_L) \cap ce(c_2, O_R)|}{|ce(c_1, O_L)|} \tag{7}$$

and

$$TR_{CSC}(O_L, O_R) = TP_{CSC}(O_R, O_L) \tag{8}$$

The $ce(ci, OJ)$ is the characteristic extract from the concept hierarchy of the ontology O_J. c_i is the concept in the similar position in both ontologies concept hierarchies. The intersection in the numerator provides the common concepts in the similar positions in both ontologies concept hierarchies.

5. Results discussions and evaluation

The learned ontology is evaluated against the reference ontology by using Lexical Precision (LP), Lexical Recall (LR), Taxonomic Precision (TP_{CSC}) and Taxonomic Recall (TR_{CSC}) measures. The obtained measures of LP, LR, TP_{CSC} and TR_{CSC} after evaluation are tabulated below in Table 9. below.

Table 9. LP, LR, TP_{CSC} and TR_{CSC} measures without concepts mapping based on Google synonyms

LP	LR	TP_{CSC}	TR_{CSC}
57.10%	28.50%	100%	100%

The LP percentage and LR percentage are increased after concepts of both ontologies are mapped based on the Google synonyms. The mapped concepts based on synonyms are {Feature is SameAs ProductAspect}, {PolarityWord isSameAs OpinionWord} and {Object isSameAs Product}. The increased LP percentage is 85.7%. A significant increase of 28.6% is observed. The increased LR percentage is 42.8%. A significant increase of 14.3% is observed.

The improved Hu and Liu (2004) opinion lexicon dataset was used in this experiment to extract the opinions from the reviews. The opinion lexicon is improved by adding 2290 unique positive words and 4800 unique negative words from SentiWordNet. Table 10. presents the details of the dataset used for this experiment.

Table 10. Opinion Lexicon details

Opinion word attributes	Values
Number of positive words	4294
Number of negative words	9582

Fifteen reviews from the five datasets from these five product categories were taken to extract the opinions and determine their orientations. The reviews are given below.

<u>Apple iphone 6s plus</u>

1. Siri is awesome do most of the work smoothly.

2. Very happy to get an iPhone 6s plus.

3. Really good phone and its my first iPhone. Great screen size and good battery life.

Letstrack Bike Tracing GPS Device

1. Good work guys. The way you guys install it, is good as it is always hidden. Now I track my bike wherever it is and I really like the parking alert feature.

2. This is best for tracking. I am using this device. This is very good for bike GPS tracking.

3. Perfect.

Micromax Canvas Tab P70221

1. Good budget tablet in 5000 range.

2. Camera is not good.

3. Amazing Tablet. wonderful wifi+3G.

Lenovo Core i5 7th Gen Laptop

1. Awesome laptop. Looks are good

2. laptop has attractive design it give much more performance than i hoped.

3. Nice product and it's able to run GTA V without lag and in high graphic.

Nikon B700 Point and Shoot Camera

1. Nice camera sensor is small but good.

2. This is the best camera in this price range.

3. Nice product with pouch and 8gb card free.

After POS tagging, the noun, adjective pairs found from the five datasets are as follows.

Iphone 6s plus: [(Siri, awesome), (iPhone 6s plus, happy), (phone, good), (screen size, great), (battery life, good)]

Letstrack Bike Tracing GPS Device: [(parking alert feature, good), (installation, good), (bike GPS tracking, good), (device, good)]

Micromax Canvas Tab P70221: [(budget, good), (camera, not good), (tablet, amazing), (wifi+3G, wonderful)]

Lenovo Core i5 7th Gen Laptop: [(laptop, awesome), (appearance, good), (design, attractive), (performance, better), (product, nice), (gaming, good), (graphics, high)]

Nikon B700 Point and Shoot Camera: [(product, nice), (camera sensor, small), (camera sensor, good), (camera, best), (price, good), (product, nice), (pouch, nice), (memory card, free)]

All the adjectives were compared with the opinion lexicon. All were identified. The identified adjectives were deemed as opinion words.

In order to determine the orientations of the opinion words, the senses of the opinion words were disambiguated by learning the context using typed dependencies (De Marneffe and Manning, 2008) and WordNet gloss. The obtained sense is used in searching for the SentiWordNet score under adjectives category. The scores were substituted in SO formula. When the obtained value after SO calculation is greater than zero, then the opinion word is termed as positive, otherwise negative. The evaluation on the orientation of the opinions using the proposed approach as compared with the baseline approaches is presented in Table 11. below.

Table 11. Accuracy (in %) of extracted opinions

Opinion Orientation Method	Pos./Neg. Adjectives	Accuracy(%)
Log-linear regression (Hatzivassiloglou and McKeown, 1997)	657/679	87.38
Orientation based on Pointwise mutual info. (Turney and Littman, 2003)	1915/2291	83.09
Lexical relations and geodesic distance (Kamps *et al*., 2004)	663 of (Hatzivassiloglou and McKeown, 2016)	88.05
Derivationally related forms based distance measurement (van Miltenburg *et al*., 2016)	310/310	74
Proposed Work	2004/4782	89.5

The results obtained in terms of accuracy with the published techniques are as shown in the Table 10, note that there are improvements in the orientation of the opinions in the work of measuring semantic orientation of adjectives using WordNet when compared with log linear classifier based semantic orientation of adjectives and PMI based semantic orientation of adjectives (Turney and Littman., 2003). However, the accuracy of the proposed method when compared to the method of van Miltenburg (2016) has increased in a significant manner.

The benefits these results bring to the E-Commerce system are that opinion orientations of the product aspects help the customers to decide which products to purchase, also to the companies in order to understand the buying behaviour of customers.

6. Conclusion and future work

The determination of opinion orientation of opinion words by using the output of learned ontology from online reviews were carried out successfully. The objective is to form all kinds of opinionated aspects from review sentences. Also the formula of SO using this sentiwordnet score has provided better opinion orientation accuracies as compared with the existing classification methods.

The accuracy obtained using the modified geodesic distance metric for SO is 89.5%. There was a significant increase of 15.5% when compared with the accuracy of the recent work on derivationally related forms based distance measurement for ascertaining SO of the opinion word. It has been observed that with the modified geodesic distance metric for SO, the opinion orientation of any opinionated word from the data corpus is evaluated when compared with the aforementioned approach. Despite high degree of accuracy is obtained with the proposed approach, there are few limitations exhibited by the approach. The proposed approach does not take into account the analysis of ironical review clauses/sentences. Also, the review containing only nouns and noun variants are learned as product aspects in the task of ontology learning. The opinion orientations of these aspects are not determined in this approach.

In future, the opinion words analysis is carried out on the detection of irony from the review sentences. Irony is context sensitive and is hard to detect. The concept hierarchies of the eVolutionary product review domain ontology are reasoned for identifying the irony context of the opinion word. Also, the opinion words with their coexisted aspects are quantified for sentiments. These sentiments together with the underlying base cases (case based reasoning) in the E-Commerce product categories database for the searched product are carefully compared. This comparison is for providing explainable sentiments based product recommendations to the E-Commerce customer.

7. Websites list

Site 1: Ginger Software (List of irregular verbs)

https://www.gingersoftware.com/content/grammar-rules/verbs/list-of-irregular-verbs/

Site 2: Speak Speak (List of irregular adverbs)

http://speakspeak.com/resources/english-grammar-rules/adjectives-adverbs/irregular-adverbs

Site 3: Enchanted Learning (List of irregular adjectives)

https://www.enchantedlearning.com/grammar/partsofspeech/adjectives/

References

Benamara F., Cesarano C., Picariello A., Recupero D. R., Subrahmanian V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *International Conference on Weblogs and Social Media*, pp. 1-7.

Buitelaar P., Cimiano P. (2008). Ontology learning and population: bridging the gap between text and knowledge. *Amsterdam: IOS, Press*, pp. 45-69. https://doi.org/10.3233/978-1-60750-072-8-187

Buitelaar P., Cimiano P., Magnini B. (2006). Ontology learning from text: methods, evaluation and applications. *Computational Linguistics*, Vol. 32, No. 4, pp. 569-572. https://doi.org/10.1162/coli.2006.32.4.569

De Marneffe M. C., Manning C. D. (2008). Stanford typed dependencies manual. *Stanford University*, pp. 338-345.

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge. MA: MIT Press.

Grondelaers S., Speelman D., Geeraerts D. (2012). Lexical variation and change. *The Oxford Handbook of Cognitive Linguistics*, pp. 988-1011. https://doi.org/10.1017/CBO9780511841613.021

Hatzivassiloglou V., McKeown K. R. (1997). Predicting the semantic orientation of adjective. *Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 174-181. https://doi.org/10.3115/976909.979640

Hu M., Liu B. (2004). Mining and summarizing customer reviews. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177. https://doi.org/10.1145/1014052.1014073

Huang H. H., Wang J. J., Chen H. H. (2017). Implicit opinion analysis: Extraction and polarity labelling. *Journal of the Association for Information Science and Technology*, Vol. 68, No. 9, pp. 2076-2087. https://doi.org/10.1002/asi.23835

Johnson M. (2002). Squibs and discussions: the DOP Estimation method is biased and inconsistent. *Computational Linguistics*, Vol. 28, No. 1, pp. 71-76. https://doi.org/10.1162/089120102317341783

Kamal A., Abulaish M. (2016). OntoLSA—An integrated text mining system for ontology learning and sentiment analysis. *Sentiment Analysis and Ontology Engineering*, pp. 399-423. https://doi.org/10.1007/978-3-319-30319-2_16

Kamps J., Marx M., Mokken R. J., De Rijke M. (2004). Using WordNet to Measure Semantic Orientations of Adjectives. *Language Resources and Evaluation Conference*, Vol. 4, pp. 1115-1118. https://doi.org/10.1.1.134.483

Lau R. Y., Lai C. C., Ma J., Li Y. (2009). Automatic domain ontology extraction for context-sensitive opinion mining. *International Conference on Information Systems*, pp. 35-53.

Liu B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Vol. 5, No. 1, pp. 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Montemagni S., Wieling M. (2016). Tracking linguistic features underlying lexical variation patterns: A case study on Tuscan dialects. *The Future of Dialects: Selected Papers from Methods in Dialectology XV*, pp. 117-134. https://doi.org/10.17169/langsci.b81.146

Morris C. (1958). The measurement of meaning. Charles E. Osgood, George J. Suci , Percy H. Tannenbaum. *American Journal of Sociology*, Vol. 63, No. 5, pp. 550-551. https://doi.org/10.1086/222316

Navigli R., Velardi P. (2010). Learning word-class lattices for definition and hypernym extraction. *Annual Meeting of the Association for Computational Linguistics*, pp. 1318-1327. https://doi.org/10.1.1.172.7430

Pang B., Lee L., Vaithyanathan S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, pp. 79-86. https://doi.org/10.3115/1118693.1118704

Polpinij J., Ghose A. K. (2008). An ontology-based sentiment classification methodology for online consumer reviews. *International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society*, Vol. 01, pp. 518-524. https://doi.org/10.1109/WIIAT.2008.68

Qiu G., Liu B., Bu J., Chen C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, Vol. 37, No. 1, pp. 9-27. https://doi.org/10.1162/coli_a_00034

Rada R., Mili H., Bicknell E., Blettner M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 19, No. 1, pp. 17-30. https://doi.org/10.1109/21.24528

Ravi K., Ravi V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-based Systems*, Vol. 89, pp. 14-46. https://doi.org/10.1016/j.knosys.2015.06.015

Ravi Kumar K. C., Teja Santosh D., Vishnu Vardhan B. (2017). Determining the Semantic Orientation of opinion words using typed dependencies for opinion word senses and Sentiwordnet scores from online product reviews. *International Journal of Knowledge and Web Intelligence*. https://doi.org/10.1504/IJKWI.2017.10010171

Santosh D. T., Vardhan B. V. (2016). PROO ontology development for learning feature specific sentiment relationship rules on reviews categorisation: A semantic data mining approach. *International Journal of Metadata, Semantics and Ontologies*, Vol. 11, No. 1, pp. 29-38. https://doi.org/10.1504/IJMSO.2016.078105

Speer R., Havasi C. (2012). Representing general relational knowledge in ConceptNet 5. *Language Resources and Evaluation Conference*, pp. 3679-3686. https://doi.org/10.1.1.383.5758

Sun C., Zuo Z. S., Lu W., Liu X. T., Guo X. L., Liu F. (2017). Visualization of the heat transfer character of dry slag discharge system. *International Journal of Heat and Technology*, Vol. 35, No. 4, pp. 793-798. https://doi.org/10.18280/ijht.350414

Tho Q. T., Hui S. C., Fong A. C. M., Cao T. H. (2006). Automatic fuzzy ontology generation for semantic web. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 6, pp. 842-856. https://doi.org/10.1109/TKDE.2006.87

Toutanova K., Manning C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Annual Meeting of the Association for Computational Linguistics*, Vol. 13, pp. 63-70. https://doi.org/10.3115/1117794.1117802

Turney P. D., Littman M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, Vol. 21, No. 4, pp. 315-346. https://doi.org/10.1145/944012.944013

van Miltenburg E. (2016). WordNet-based similarity metrics for adjectives. *Global WordNet Conference, Global WordNet Association*, pp. 414-418.

Volkova S., Wilson T., Yarowsky D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. *Conference on Empirical Methods in Natural Language Processing*, pp. 1815-1827.

Wang T., Hirst G. (2012). Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering*, Vol. 18, No. 3, pp. 313-342. https://doi.org/10.1017/S1351324911000210